Thanks to the reviewers for their effort towards providing feedback on our paper.

Reviewers 1 and 2 – iterate averaging for ResNet on CIFAR-10: We will be happy to perform experiments evaluating the performance of iterate averaging for CIFAR-10 with a residual network and incorporate these results in the final version. These experiments are somewhat subtle since, performing Polyak averaging right from the beginning might not be good due to the non convex nature of the problem. One needs to find the right number of iterations after which we need to begin Polyak averaging. These experiments (on the performance of iterate averaging) could be of great help to the community since iterate averaging is frequently discussed but no thorough experiments results have been published to the best of our knowledge.

We will now address specific reviewer comments.

Reviewer 1: Thank you for your comments.

Moving CIFAR-10 experiments to the start of the paper: The suggestion on presenting the experiments section in the introduction is quite interesting. We will think about it carefully before the final version. Even if the CIFAR-10 experimental results are not strictly novel, our primary goal was to perform a thorough grid search based experiment in order to understand what is the behavior of the final iterate with both polynomial and exponentially decaying step size schedules. To the best of our knowledge, we are unaware of experiments that perform these experiments thoroughly and with reasonably large neural networks.

Minor comments : will address 1,2,3. Note that for 3, the x-axis is in log scale (instead of y-axis).

Reviewer 2: Thank you for your comments.

Experiments on least squares objective: We will present results on the least squares objective with both iterate averaging as well as polynomially and exponentially decaying step sizes. We also note that in figure 1 (right) of the paper, we present results with grid searches on optimizing a quadratic objective - see caption of figure 1 as well as appendix E.1 in the supplementary section. The (bad) performance of polynomially decaying learning schemes, as well as the (good) performance of exponentially decaying step sizes for the quadratic objective implies that these perform similarly for the least squares objective - this is through the result of lemmas 8-11 in the appendix.

What problems do these lower bounds hold for? Our construction relies on objective functions that have bad condition numbers, and that is rather typical for many machine learning problems. Furthermore, note that going beyond least squares, for objectives that satisfy notions of local quadratic approximation (for e.g. self-concordance), our results (after going through some more formal arguments) has the potential to be made to apply towards the rate near the optimum. We will include this discussion in the final version.

Reviewer 3: Thank you for your review.